COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Huiwen Chang

Lecture #7
February 25, 2014

---

1 A Lower Bound on Sample Complexity

In the last lecture, we proved an upper bound about how many examples are needed for PAC learning involving the VC dimension. Then we started talking about that the VC dimension also provides a lower bound for learning. While the upper bound gives the sufficient condition for PAC learning, the lower bound gives the necessary condition which says if the examples is too small, then the concept is not PAC learnable.

Last time, we gave a false proof of the lower bound. We generate a random training set, and then choose a concept from $\mathcal{C}$ which labels exactly the opposite to the prediction given by our hypothesis. This is cheating because the concept has to be chosen before the training set is generated. In the following, we give a correct proof of the lower bound:

Theorem 1. *Let $d = VC\text{-}dim(\mathcal{C})$. For any algorithm A, there exists distribution $\mathcal{D}$ and a concept $c \in \mathcal{C}$, such that if A is given sample $\mathcal{S}$ of $m \leq d/2$ examples, then*

$$Pr[err(h_A) > \frac{1}{8}] \geq \frac{1}{8}$$

*An alternative(rougher) statement: if $\epsilon \leq 1/8$ and $\delta \leq 1/8$ then we need more than $d/2$ examples for PAC learning.

Proof. According to the definition of the VC dimension, we know that there exist d points $z_1, ..., z_d$ shattered by $\mathcal{C}$. Let D be uniform over $z_1, ..., z_d$. Let $\mathcal{C}' \subseteq \mathcal{C}$ have one representative for every labeling of the shattered points $z_1, ..., z_d$. Then we know $|\mathcal{C}'| = 2^d$. Then we choose c uniformly at random from $\mathcal{C}'$.

Let's think about two experiments about how to generate variables:

- Experiment 1:
  c is chosen uniformly at random from $\mathcal{C}'$.
  $\mathcal{S}$ is chosen at random (according to D) and labeled by c.
  h_A is computed from $\mathcal{S}$.
  The test point x is chosen (from D).
  We try to measure: $Pr[h_A(x) \neq c(x)]$.

- Experiment 2:
  Unlabeled part of $\mathcal{S}$ is chosen.
  Random labels $c(x_i)$ are chosen for $x_i \in \mathcal{S}$.
  h_A is computed from labeled $\mathcal{S}$.
  The test point x is chosen.
  If $x \notin \mathcal{S}$ then label $c(x)$ is chosen uniformly at random.
  We try to measure: $Pr[h_A(x) \neq c(x)]$.

Though the order is flipped in the two experiments above, we claim that they produce the same distribution of random variables and same probability measure. This is because the unlabeled sample S is generated independently of the choice of the labels, and the label for x is also chosen independently of the samples S, labels of other points, and the prediction of hypotheses. So in both experiments, the probability is given over random variables concept c, sample S, and the test point x. We denote it as $Pr_{c,S,x}[h_A(x) \neq c(x)]$. Let's work on experiment 2, and we have

$$\begin{aligned}
& Pr_{c,S,x}[h_A(x) \neq c(x)] \\
\geq\ & Pr_{c,S,x}[x \notin \mathcal{S} \wedge h_A(x) \neq c(x)] \\
=\ & \underbrace{Pr_{c,S,x}[x \notin \mathcal{S}]}_{\geq 1/2 \text{ because } m \leq d/2 \text{ and } x \text{ is uniform chosen}} \cdot \underbrace{Pr_{c,S,x}[h_A(x) \neq c(x) | x \notin \mathcal{S}]}_{=1/2 \text{ because } c \text{ is a random guess}} \\
\geq\ & \frac{1}{4}
\end{aligned}$$

According to marginalization $Pr[a] = E_X[Pr[a|X]]$, we have

$$Pr_{c,S,x}[h_A(x) \neq c(x)] = E_c[Pr_{S,x}[h_A(x) \neq c(x)]]$$

By the fact that if $E[X] \geq b$, then there exists $x \in X$ such that $x \geq b$, we can know there exists $c \in \mathcal{C}' \subseteq \mathcal{C}$ such that

$$Pr_{S,x}[h_A(x) \neq c(x)] \geq \frac{1}{4}$$

Using marginalization again, we can get

$$\begin{aligned}
\frac{1}{4} &\leq Pr_{S,x}[h_A(x) \neq c(x)] = E_S[Pr_x[h_A(x) \neq c(x)]] \\
&= E_S[err(h_A)] \\
&\leq Pr_S[err(h_A) \leq \frac{1}{8}] \cdot \frac{1}{8} + Pr_S[err(h_A) > \frac{1}{8}] \qquad (1) \\
&\leq \frac{1}{8} + Pr_S[err(h_A) > \frac{1}{8}]
\end{aligned}$$

The inequality (1) comes as follows: for $X \in [0, 1]$,

$$\begin{aligned}
E[X] &= \sum_{x \in X} Pr(x) \cdot x \\
&= \underbrace{\sum_{x: x \leq 1/8} Pr(x) \cdot \underbrace{x}_{\leq 1/8}}_{Pr[X \leq 1/8]} + \underbrace{\sum_{x: x > 1/8} Pr(x) \cdot \underbrace{x}_{\leq 1}}_{Pr[X > 1/8]} \\
&\leq Pr[X \leq \frac{1}{8}] \cdot \frac{1}{8} + Pr[X > \frac{1}{8}]
\end{aligned}$$

□

2 Inconsistent Model Hypotheses

So far we have only dealt with the situation in which the hypotheses is consistent, and we focused on the samples needed for learning in that space, but what to do if you cannot find a consistent hypotheses? There are several reasons it may not be consistent as follows:

- The concept $c \notin \mathcal{H}$;($\mathcal{H}$ is not powerful enough to represent the truth.)
- $c \in \mathcal{H}$, but it's just a too hard computational problem to find it;
- c may not exist. (We always assume that there's a target concept c that is a functional mapping that maps each instance to a label, but the reality is not that case. Consider weather prediction—the forecaster only estimates and reports the probability of snowing tomorrow. We believe there is an intrinsic randomness, or say it's too hard to *model* in a deterministic form by requiring so much knowledge.)

So now we work with a more realistic model where there might not exist the functional relationship. We can generalize our usual model: We assume we have examples (x, y) where $x \in X, y \in \{0, 1\}$. Now we let (x, y) be random according to distribution D on $X \times \{0, 1\}$. (Unlike our earlier model, the label y here is random). It follows from the chain rule that:

$$Pr_D[(x, y)] = Pr_D[x] \cdot Pr_D[y|x]$$

We can think the process as x being first generated by $Pr_D[x]$ and then y being generated according to its conditional distribution $Pr_D[y|x]$. In the PAC model where the label is deterministic, $Pr[y|x]$ is either 0 or 1, while in this new model, it's between 0 and 1. We also need to modify the generalization error from $err_D(h) = Pr_{x\sim D}[h(x) \neq c(x)]$ to

$$err_D(h) = Pr_{(x,y)\sim D}[h(x) \neq y]$$

Now, the first question is that, "With complete knowledge of distribution D, how small can the generalization error be?" Let's start with a simpler problem — tossing a coin with a known bias. The coin comes up heads with probability p. In this case, to minimize the probability of an incorrect prediction, our strategy is

$$\begin{cases} Head, & p > \frac{1}{2}; \\ Tail, & p < \frac{1}{2}; \\ arbitrary, & p = \frac{1}{2}. \end{cases}$$

Consider each x has a coin to flip, so the optimal decision rule is similar with before:

$$h_{opt}(x) = \begin{cases} 1, & Pr_D[y = 1|x] > \frac{1}{2}; \\ 0, & Pr_D[y = 1|x] < \frac{1}{2}; \end{cases}$$

The optimal prediction rule is called the "*Bayes Optimal Classifier*" or "*Bayes optimal decision rule*", and the optimal possible error $err_D(h_{opt}) = \min_h err_D(h)$ is called the "*Bayes error*". It provides a lower bound on the error over all hypotheses regardless of computational power.

Now, our goal is to minimize $err_D(h)$ over $h \in H$. We then introduce a natural approach: Given examples $(x_1, y_1), ..., (x_m, y_m)$ chosen independently at random from D, we try to minimize the *training error* with indicator variables(1 if $h(x_i) \neq y_i$):

$$\widehat{err}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x_i) \neq y_i\} \tag{2}$$

So suppose you could find $\widehat{h} = arg\min_{h\in H} \widehat{err}(h)$. We also suppose you could show that, with probability $1 - \delta$, for any $h \in \mathcal{H}$,

$$|\widehat{err}(h) - err_D(h)| \leq \epsilon \tag{3}$$

Then for all $h \in H$:

$$\begin{aligned} err_D(\widehat{h}) &\leq \widehat{err}(\widehat{h}) + \epsilon \\ &\leq \widehat{err}(h) + \epsilon \\ &\leq err_D(h) + 2\epsilon \end{aligned}$$

Therefore, the hypotheses $\widehat{h}$ will have a generalization error close to the lower bound of the error for all hypotheses in $\mathcal{H}$:

$$err_D(\widehat{h}) \leq \min_{h \in H} err_D(h) + 2\epsilon$$

But, this approach has two things to deal with:

- The computational problem about how to minimize the training error in (2);
- The statistical problem in (3) which implies the training error is a good approximation of true error for all hypotheses in $\mathcal{H}$.

The bound in (3) is also called a "*uniform convergence bound*". We also name $err(h)$ as *true/generalization error* or *true risk*, $\widehat{err}(h)$ as *training/empirical error* or *empirical risk*, and the approach of minimizing the training error as *empirical risk minimization*.

In order to prove a uniform convergence bound, we first move to a more abstract setting. Define random variables $X_1, ...X_m$, i.i.d $X_i \in [0, 1]$ for all $i = 1, \ldots, m$. Let $p = E[X_i]$, $\widehat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i$. In fact, if we denote X_i to be $1\{h(x_i) \neq y_i\}$, we can see $\widehat{p}$ is the training error and p is the generalization error. We want to show how close $\widehat{p}$ is with respect to the mean p.

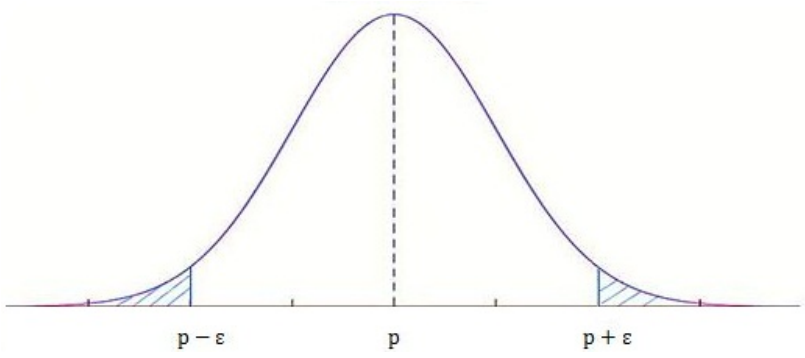

Figure 1: Illustration of concentration inequality or tail bound on $\widehat{p}$

Let's look at the distribution of $\widehat{p}$ in Figure 1, next time, we will show the tail $Pr[\widehat{p} > p + \varepsilon]$ and $Pr[\widehat{p} < p - \varepsilon]$ are really small, which is called "tail bounds" or "concentration inequalities". In the next lecture, we will provide a proof of a general bound – the Chernoff bound.